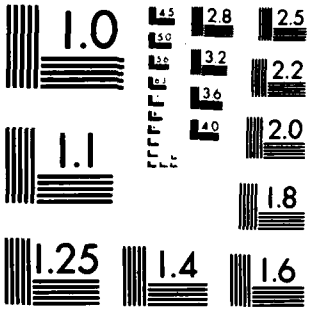
AD-A137 936 ON THE LOGICAL DEVELOPMENT OF STATISTICAL MODELS(U) 1/1
WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER
D PENA DEC 83 MRC-TSR-2608 DAAG29-80-C-0041

UNCLASSIFIED F/G 12/1 NL

END
DATE FILMED
3 84
DTIC



MRC Technical Summary Report #2608

ON THE LOGICAL DEVELOPMENT OF STATISTICAL MODELS

Daniel Peña

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

December 1983

(Received September 15, 1983)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

ON THE LOGICAL DEVELOPMENT OF STATISTICAL MODELS

Daniel Peña*

Technical Summary Report #2608
December 1983

ABSTRACT

This paper presents a classification of statistical models using a simple and logical framework. Some remarks are made about the historical appearance of each type of model and the practical problems that motivated them. It is argued that the current stages of the statistical methodology for model building have arisen in response to the needs of more sophisticated procedures for building dynamic-explicative types of models. Some potentially important topics for future research are included.

AMS (MOS) Subject Classifications: 62-03, 01A99, 62A99

Key Words: statistical models, methodology, history of statistics, robustness

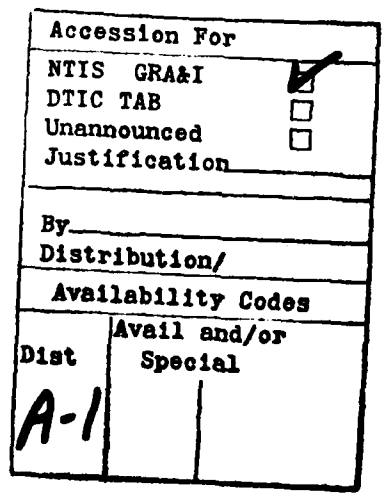
Work Unit Number 4 (Statistics and Probability)

*Statistics Department, ETSII, Universidad Politecnica de Madrid, Spain

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the United States-Spanish Joint Committee for Educational and Cultural Affairs.

SIGNIFICANCE AND EXPLANATION

It is shown in this paper how statistical models can be classified in a simple framework according to the type of data (extrapolative versus explicative models) and the a priori knowledge of the variables (static versus dynamic situations). This classification throws light on some polemical historical points in the history of statistics, is useful in finding the mainstream of statistical thought, and allows a meaningful interpretation of the evolution of statistical methodology as a response to the needs of those models. Finally, this analysis has suggested many topics that appear to be promising for future research.



The responsibility for the wording and views expressed in this descripive summary lies with MRC, and not with the author of this report.

ON THE LOGICAL DEVELOPMENT OF STATISTICAL MODELS

Daniel Peña*

1. INTRODUCTION

Anyone looking through a library of statistical books will find that this discipline seems to be divided into many different branches that sometimes appear to be loosely related. There are texts on data analysis and stochastic processes, on nonparametric statistics and decision theory, on the linear model and categorical data. But the integrated relationship between these subjects is not easy to find. Subject classifications of statistical subjects, such as the AMS (MOS), or even the usual classification of interests to the members of statistical societies, are not much more helpful. Where does the unity of statistics lie?

It has been argued (see for instance Box, Hunter and Hunter (1978)) that the unity of statistics as a science is based on the general goal of building mathematical models for non-deterministic systems to understand them and/or to make forecasts or decisions. From this point of view, the strength of statistics lies in providing scientists with a general methodology to learn from reality and to approximate sequentially their scientific goals in a coherent and systematic way.

This paper explores the relationship between statistical models and the methodology that has been developed to build them. First, a classification of statistical models is presented that includes those models which constitute the backbone of this science. Second, their historical evolution is briefly revised to show that the development of these models flows parallel to the

*Statistics Department, ETSII, Universidad Politecnica de Madrid, Spain

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the United States-Spanish Joint Committee for Educational and Cultural Affairs.

growth of statistics as a discipline. Third, it is argued that the appearance of a new class of models has led to changes in methodological procedures, and the current statistical methodology can be better understood in this context. Finally, some concluding remarks will be made.

2. A CLASSIFICATION OF STATISTICAL MODELS

Any sensible statistical model is built using two kinds of resources: data and logical reasoning. The former introduces the inductive aspect of inference and the latter the deductive part. Assuming that the objective of a statistical investigation is to study a (possibly vector) random variable y, a useful classification of statistical models can be obtained according to these two dimensions. We shall call extrapolative the class of models that are built using only past values of the variable y, and explicative those which also take into account the values of other explicative variables x. In either case, logical reasoning and *a priori* knowledge about the variable y and the way its values are observed will lead us to either a static or a dynamic model.

Table 1 shows a classification of statistical models according to these two basic criteria. In all cases the model can be expressed, perhaps after some transformation to obtain an additive decomposition, as:

$$\text{VARIABLE} = \text{SYSTEMATIC VALUE} + \text{NOISE} \qquad (1)$$

An extrapolative-static model, here after termed a type I model, can be characterized as follows: (1) The systematic value μ is a constant and (2) the noise has a probability distribution which depends on a vector of parameters $\underset{\sim}{\theta}_2$. Type I models include scalar and vectorial probability distributions. Usually, the noise has an expected value equal to zero, so that μ is the mean value of y. When y is a discrete variable the value μ often appears as a parameter in the distribution of the noise.

A more comprehensive approach to explain the behavior of y is to decompose μ into two terms. The first contains the general level of y and the second, $f_2(\underset{\sim}{x},\beta)$, the effect of a known vector $\underset{\sim}{x}$ of predictor variables. In this way we obtain an explicative-static or type II model and

Table 1

Classification of Statistical Models

	STATIC	DYNAMIC
EXTRAPOLATIVE	Model: $y_i = \mu_1 + u_i; \quad u_i \sim F(\underline{\theta}_2)$ - *Scalar probability distribution* - Vectorial probability distribution TYPE I	Model: $y_t = \mu_t + u_t; \quad \mu_t = f_1(\mu_{t-1}, \mu_{t-2}, \ldots, \underline{\gamma}) + \varepsilon_t$ $u_t \sim F_t(\underline{\theta})$ - *Scalar Stochastic Processes* - Vectorial Stochastic Processes TYPE III
EXPLICATIVE	Model: $y_i = \mu + f_2(\underline{x}, \underline{\beta}) + u_i$ $u_i \sim IN(0, \sigma^2)$ - Linear models (scalar/vectorial) - Log-linear and categorical explicative models TYPE II	Model: $y_t = \mu_t + f_3(\underline{x}_t, \underline{\beta}) + u_t$ $u_t = f_4(y_{t-1}, y_{t-2}, \cdot; \underline{\phi})$ $u_t \sim IN(0, \sigma^2)$ - Dynamic Transfer Function Models - Multivariate Transfer Function Models TYPE IV

it is often assumed, citing the central limit theorem, that the noise has a normal distribution. The most important class of type II models is the linear model, obtained when f_2 is a linear function of the unknown parameters $\underline{\beta}$:

$$f_2(\underline{x},\underline{\beta}) = x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k \ .$$

If the variable y is qualitative, it is possible to write a similar expression (after some transformation) and we obtain the class of log-linear and categorical explicative models.

The dynamic counterparts of these models are designed to take into account any sequential character of the observations. There are two basic approaches to represent the dynamic structure of a system: The first approach uses a representation that is similar to the static model, but allows the parameters to change, stochastically, over time. This is generally called the state-space representation of a dynamic variable. The second approach relates current values of the variable to past values using a difference equation representation with parameters that are assumed fixed. For type III models (extrapolative-dynamic) Table 1 shows the state-space approach which is characterized by two properties: (1) The systematic part, or expected value, of the variable changes over time according to a given structure f_1 which is typically assumed to be linear, and (2) the noise has a probability distribution that may or not change over time. For discrete variables, the parameters of this probability distribution often depend on μ_t.

For example, the simplest representation for a discrete stochastic variable with two possible states and first order Markov dependency is:

$$\begin{aligned} &y_t = \mu_t + u_t \ . \\ &p(u_t = 1 - \mu_t) = \mu_t, \quad p(u_t = -\mu_t) = 1 - \mu_t \ . \\ &\mu_t = p_{01} + \mu_{t-1}(p_{11} - p_{01}) \end{aligned} \qquad (2.1)$$

where y_t can only have the values zero and one, μ_t is its expected value

(in this case the probability that y_t equals one), and p_{01} and p_{11} represent the transition probabilities between states zero and one. This is the classic two state Markov Chain. As we see the probability distribution of the noise is changing across the time.

As another example, a classical state-space representation of a simple time series model is:

$$\begin{aligned} y_t &= \mu_t + u_t \\ \mu_t &= \phi\mu_{t-1} + \varepsilon_t \qquad (2.2) \\ u_t &\text{ and } \varepsilon_t \text{ are independent normal variables} \end{aligned}$$

Both models are particular cases of the general representation of Table 1.

The second approach to the modeling of stochastic processes is the difference equation representation which has the general structure:

$$\begin{aligned} y_t &= k + g(y_{t-1}, y_{t-2}, \ldots, \underline{\phi}) + u_t \qquad (2.3) \\ u_t &\sim F_t(\underline{\theta}) \ . \end{aligned}$$

Here the explicit dependence of y_t on its past values is displayed in the structural equation. This approach has been particularly useful in time series models. For example, model (2.2) could be written using the backshift operator, B, as:

$$y_t - \phi y_{t-1} = \varepsilon_t + (1 - \phi B)u_t$$

or, since the addition of white noise plus an MA(1) process on the right-hand side is itself an MA(1) process:

$$(1 - \phi B)y_t = (1 - \theta B)a_t$$

which is an ARMA (1,1) process with standard (2.3) representation:

$$y_t = \sum_{i=1}^{\infty} \pi_i y_{t-i} + a_t$$

or

$$\pi(B)y_t = a_t$$

where $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 \ldots$ and its coefficients can be found using the relationship:

$$(1 - \phi B)(1 - \theta B)^{-1} = \pi(B)$$

The generalization of type III models to include exogenous variables is now straightforward and is shown in Table 1. Again two kinds of representation are possible in the same spirit as model III: a difference equation representation, which is assumed in Table 1, and a state-space approach. As both ways can be considered equivalent from a mathematical point of view, the choice between them has to be made by methodological considerations: which of these makes the identification of the process simpler? which is better for estimation and diagnostic checks? (See Peña (1978) for a discussion of the advantages of the difference equation representation).

In both cases, the functions f_3 and f_4 are typically assumed to be linear:

$$f_4(y_{t-1}, \ldots, \underline{\phi}) = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots ,$$

$$f_3(\underline{x}_t, \underline{\beta}) = \nu_1(B)x_{1t} + \ldots + \nu_k(B)x_{kt}$$

with

$$\nu_i(B) = \beta_{0i} + \beta_{1i}B + \beta_{2i}B^2 + \ldots .$$

Type IV models include transfer function or dynamic regression representations, intervention models and multivariate time series with exogenous variables, among others. It is interesting to note that so far little has yet been done to develop this class of models for the case in which y is a qualitative variable, but it seems reasonable to hope that it will be studied in the near future.

Type I and type II static models can be considered to constitute the "first generation" of statistical models. They are still the most frequently

applied models and their study is the core of the statistical curriculum in most universities. Many textbooks even identify statistical model building with this class.

Type III and IV models, the dynamic counterpart of the above, may be seen as the "second generation". They have emerged in statistical practice in our century and they are still taught infrequently and, considering that we are living in a dynamic world, less often applied. However, much research has been done in recent years on dynamic models and widespread use of them in the years to come can be foreseen.

The new trends in statistical modeling could be characterized as assuming more generality in the model structure. The classical representations most always assume, first, the same pattern or distribution for all the observations; second, linearity in the response; and third, a simple distribution for the noise. Although the first assumption was relaxed for Type I models in the XIX century to accomodate outliers in the sample, its extension to Type II models has been made only in the last twenty years and its application to dynamic models is still just beginning. The same could be said of nonlinear statistical models and of the search for more general noise structures. Table 2 summarizes some of these possible generalizations for each type of model.

Table 2

Generalization of Classical Models

TYPE I : $y_i = \mu_i + u_i$; $u_i \sim F_i(\underline{\theta})$

- μ_i is not the same for all observations
- u_i does not follow the same distribution
  - heavy-tailed
  - contaminated model
  - heteroscedastic

TYPE II : $y_i = \mu_i + f_i(\underline{x},\beta) + u_i$; $u_i \sim F_i(\underline{\theta})$

- μ_i is not the same for all observations
- f_i is not linear
- $f_i(\underline{x},\beta)$ does not include the same variables
- The parameters β are not the same
- The distribution F is not the same

TYPE III : $y_t = f_1(y_{t-1},\ldots,\phi) + u_t$; $u_t \sim F_t(\underline{\theta}_t)$

- The function f is nonlinear
- The parameters of f are changing
- The functional form of f is changing
- The distribution of u_t does not follow the same pattern for all the variables

TYPE IV : $y_t = f_4(y_{t-1},\ldots,\phi) + f_3(\underline{x}_t) + u_t$

- The functions are nonlinear
- The parameters of f_4 and/or its form are changing
- The parameters of f_3 and/or its form are changing
- The distribution of u_t does not follow the same pattern for all observations

3. SOME COMMENTS ON THE HISTORY OF MODELS

Where does the history of statistics begin? If we think of statistics as the science which studies how to obtain information about reality by means of models, it is clear that we should take the emergence of the first statistical model as the natural birth of this discipline. According to the previous section we should look for the appearance of the simplest Type I model: A static extrapolative model for a count variable with only two possible different values. Let us call this variable y and agree to establish $y = 1$ and $y = 0$ to represent these two values. The model would be:

$$y = p + u \tag{2.1}$$

where p is some constant which represents the level or mean value of y. If p is the expected value, the probability distribution of the noise u is completely determined and should be:

$$\begin{aligned} u &= 1 - p \quad \text{with probability} \quad p \\ u &= -p \quad\quad \text{with probability} \quad 1 - p \ . \end{aligned} \tag{2.2}$$

This simple model has some relevant features. First, the probability distribution of the noise depends only on the constant p which is therefore not only the expected value of the variable but also the parameter that specifies the probability distribution of the noise. Second, this model could be used as a basic block to formulate more complex types of models and, in particular, the straightforward extension:

$$f_n(A) = p_A + \varepsilon$$

where $f_n(A)$ is the relative frequency of some event, p_A its probability and ε the noise which has a binomial distribution with parameters n, number of trials, and p.

The fact that p was both a probability and an expected value explains why any of these ideas could be taken as a starting point for the probability

calculus, (see De Finetti (1975) for a subjective development of this aspect) and explains why both ideas were so closely related in the XVII century. Third, as is well known, this model was initially applied to games of chance, where the value p did not need to be estimated but could be conjectured from symmetry considerations, and so the estimation problem did not arise.

The appearance of this model required not only the knowledge that relative frequencies stabilize in the long run, but also the concept of a probability distribution for the noise. Although the first point was recognized by writers of the Renaissance at the end of the fifteenth century, we had to wait until the XVII century for the emergence of the concept of expectation (Huygens in 1657, see Maistrov (1974), p. 49) and of the idea of a probability distribution as a mathematical model to be applied to a large class of problems. (J. Bernoulli in "Ars Conjectandi", see Maistrov (1974), pp. 68-69 and also Todhunter (1865))

The next important step occurred with the development of a statistic-extrapolative model for a continuous variable. The practical problem connected with it was the modeling of astronomical measurement errors. Following the publication of Newton's theory, many leading mathematicians and astronomers of his century were concerned with contrasting the theory with existing data. It soon became clear that a theory was necessary to handle the several slightly discrepant observations of the same quantity that were obtained in repetitive measurement, and this led to the model:

$$y = \mu + u$$

where now μ cannot be calculated by logical reasoning about the symmetry of a die or some other gaming device, but must be estimated from the data. This problem led Daniel Bernoulli to the discovery of the maximum likelihood method of estimation. (See D. Bernoulli (1777))

The next logical stage in this evolution was the emergence of Type II models. Laplace stated at the beginning of the nineteenth century that the variability of astronomical variables could be explained by taking into account other measurement variables x. He assumed a linear relationship and was able to estimate the parameters of the functional equation by minimization of the absolute errors. (See Stigler (1975)). In the application of Type I and II static models to measurement data, it made sense to think of the error component as explained by a large number of small additive effects, which led, via the central limit theorem, to the normal curve.

If the Type I model brought up the central problem of the estimation of model parameters, the use of Type II model accounted for the development of the least squares method of estimation by Legendre and Gauss (see Seal (1967) for a good history of this problem) and for the general development of the linear model by Gauss.

An interesting fact that emphasizes the importance of practice in the development of statistical models, as stressed by Box (1983), is the following. As we have seen, Type I models were first developed for qualitative variables. However, Type II models for quantitative variables have not been fully developed until recently (with the log-linear model and related topics) when the need of a better understanding of complex sample surveys had become acute.

From the middle of the XIX century, the theory of Darwin became the moving force for the development of statistical thought, replacing Newton's theory in leading the mainstream of statistical advances.

The application of Type I models to biological data showed the need for new kinds of probability distributions to cope with the highly skewed and heavy-tailed distributions that were observed. Karl Pearson enlarged the

class of statistical models by proposing a system of frequency distributions, and faced the problem of estimation by introducing the method of moments.

Fisher was the first to imagine the linear model with qualitative variables, and to analyze it he introduced the Analysis of Variance. The importance of Fisher in the development of Type II statistical models has been clearly discussed by Box (1978).

The concept of stochastic dependence among variables and the need to build dynamic models did not appear until the end of the nineteenth century, and so it can be safely stated that dynamic models belong to the present century, although the roots of a stochastic process can be found in the early problem of the duration of play, a situation that can be regarded as a linear random walk with absorbing barriers (see Thatcher (1957)). Type III stochastic models first emerged in the study of the extinction of surnames by Galton and Watson, which led to branching processes (see Kendall (1966) and Harris(1963)), the work of Bachelier and Poincaré on the random walk to explain the Paris stock market, the work of Einstein on Brownian motion (see Brush (1968)), and the seminal work of Markov in 1907 on stochastic chains (see Maistrov (1977)). It should be noted that much remains to be done to unify the several varieties of Type III models that have since been developed.

The state space representation of stochastic processes has been mainly used in the control theory and engineering literature. See Ephremides and Thomas (1973) and Ephremides (1975) for a review of some benchmark contributions. For continuous variables observed as a time series, the difference equation representation has been in use since Yule (1927). It is interesting to note that here, as in other areas of statistics, the parametric models, such as those advocated by Box and Jenkins (1970), have shown clear

advantages over the non-parametric approaches, such as those developed for spectral analysis.

Although some kinds of Type IV models have been discussed in the statistical and econometric literature, only in the last few years have there been important steps toward linking this class of model with general statistical methodology. However, there is still no complete, coherent theory for building these kinds of dynamic models for qualitative variables.

As might be expected, the vectorial representation of all these types of models lagged behind the scaler forms. The first multivariate distribution did not arise until the middle of the XIX century (see Lancaster (1977)) and the multivariate linear model was first studied in the 1930's and 1940's. The study of vector representations of dynamic models is still far from complete. See Karlin and Taylor (1975), Hannan (1970) and Tiao and Box (1981).

As far as the generalizations of these models are concerned, the concept of setting up a model in which all the observations do not follow the same distribution emerged in the middle of the XIX century for Type I models. Glaisher (1872-73) assumed that the data were normally distributed with a common mean k but with unknown and unequal variances (see Barnett and Lewis (1978)). Since then, the outliers problem has led to new structures for the noise that have gone in two directions. The first, and the most rewarding one, is to embed the classical noise structure into a new, more general, distribution. This is the path followed by Jeffreys (1932) for the contaminated normal model, by Tukey (1960) for the mixture of distributions with different variances, by Box and Tiao (1973) for the exponential family of heavy-tailed distributions or by Box and Cox (1964) for the transformation problem. The second path is to assume a change in the systematic part that,

for Type I models, leads to slippage models (see Barnett and Lewis (1978)). The extension of these ideas to Type II models has been partially made in the last twenty years but there are still very few results for dynamic models.

The appearance of computers and of nonlinear optimization techniques has made possible the growing interest in nonlinear models. Broadly speaking, changes in the systematic part could be considered as a special kind of nonlinear structure, but again a general structure is still lacking.

4. MODELS AND METHODOLOGY

The methodology of statistics has incorporated new tools and procedures as the need to build new classes of statistical models has appeared. The application of Type I models for continuous variables motivated the problem of estimation. The analysis of residuals and the need for careful diagnostic model checking arose in the development of the linear model. The model identification stage was first clearly advocated for building ARIMA time series models. Finally, the development of more complex Type IV dynamic models is showing the need for a new stage in which the sensitivity of the model to the data is explored.

If we reviewed the text-books of the 40's and 50's and even many from the 60's, it would be clear that the core of these books refers to Type I kinds of models. In this context, the main problem of statistical model building was considered to be the estimation and hypothesis testing problems, and for many authors the concepts of statistical inference, statistical model building, estimation of parameters and hypothesis testing were considered synonymous. The methodology advocated was therefore static. It was assumed that the statistician decided from the outset what kind of probability distribution should be adequate to the situation according to his "a priori" knowledge and then he either went through the traditional process of interval estimation and hypothesis testing or, he used a Bayesian approach to estimate the parameters of the model. In the classical framework, a goodness of fit test of the distribution could be made to confirm the adequacy of the assumed model.

Those textbooks rarely suggested that the initial assumptions about the distribution could be wrong, and the methodology so stated was static although full of mathematical harmony. When models of Type II were built, the above approach seemed obviously inadequate, but the kind of iterative process needed

for any successful application of these models was regarded by most authors as, somehow, "cheating with data" and so not deserving of a place in scientific statistical textbooks. Besides, the fitting of a multiple regression model was so cumbersome from a computational point of view that the estimation of the parameters of any model became the crucial problem.

The computer made it possible to integrate the Type II models into common statistical practice. Statisticians soon became aware of the need to apply diagnostic checks to the residuals of a linear model and to use these checks to reformulate it and to learn from its deficiencies.

Although the analysis of residuals had been done informally before, in one way or another, by all good statisticians, the systematic study of how to identify departures from underlying assumptions and the need to integrate this knowledge into the model building process did not arise until the 60's. Anscombe and Tukey (1963) and Draper and Smith (1966) were, among many others, leaders of this movement, and their work has had a strong influence in establishing diagnostic checking of the model as an important part of statistical methodology for building models.

It gradually became clear that the same graphical displays and informal analysis that were useful to check residuals could be used earlier in the model building process to identify possible alternative model structures. The need for these tools was especially urgent for dynamic models in which the lag relationship is normally unknown and cannot be obtained from a priori reasons. Box and Jenkins (1970) advocated the need for an identification stage as a fundamental step in the statisitcal model building methodology. Their work on time series models made clear how to investigate empirically the functional form of μ_t for Type III models, where we cannot rely on external information to do so, as was supposed (often wrongly) for Type II models. In

fact, there is one important difference between a classical linear model and an ARIMA model: in the former, the function f of Table 1 is either completely known (as in a designed experiment) or is somehow controlled by the statistician through the choice of the explicative variables, whereas in the latter the structure of μ_t is unknown and should be determined from the experimental data.

An important aspect of statistical model building philosophy is the concept of robustness. In addition to criterion robustness and inference robustness (see Box and Tiao (1973)), it is important to take into account the "data" robustness. Data can be thought of as the ground on which we build the model structure, and the degree to which alternative models rest on the data can be quite different. The point goes further than the need for procedures for outlier rejection because the question we should ask is to what extent the basic properties of the model are due to a small fraction of data values. This stage can be called the analysis of data sensitivity or data robustness. To accomplish this task, techniques such as cross-validation and influential observations in the spirit of Cook and Weisberg (1982) and Belsley, Kuh and Welsch (1980) can be applied, although much remains to be done in this field. The importance of these works is to point out that even in the well-known and extensively studied linear model, the effects of small subsets of data on the properties of the model can be unexpected. Needless to say, when we build a complex Type IV model it is of outstanding importance to find out if its main properties are based on a small subset of the data to prevent us from building a whole theory on a small amount of information.

Figure 1 displays this stage in the framework of statistical methodology. The figure shows together the logical steps of the iterative model building philosophy and the parts of statistical knowledge that are

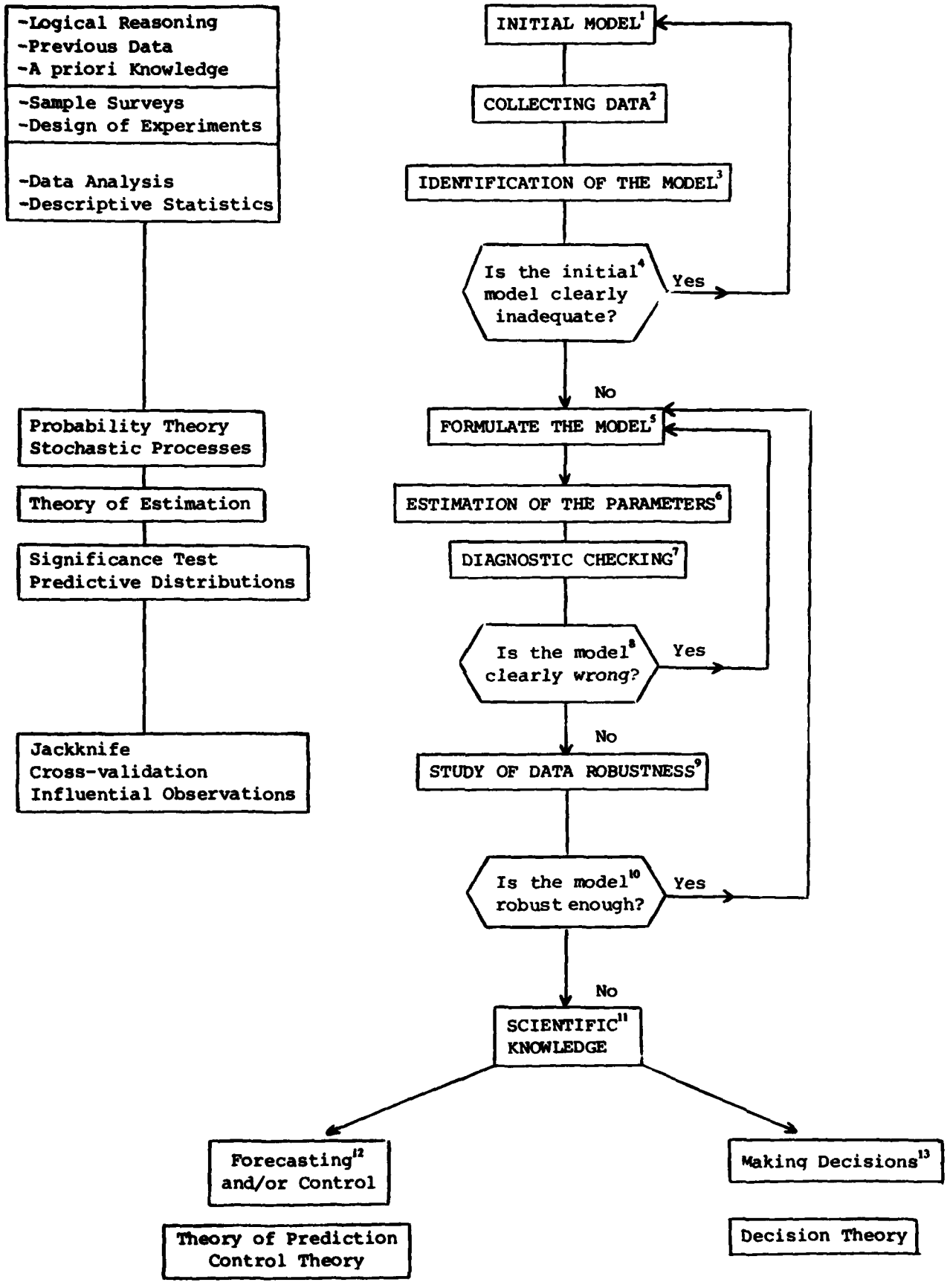


Figure 1. Statistical Methodology

adequate for each purpose. Data Analysis can be seen as the set of procedures to identify logical patterns in data that can lead to entertain an initial model. The study of data sensitivity or data robustness has been differentiated in the figure from the stage of diagnostic checking. The objective of diagnostic checks is to investigate if any of the assumed hypothesis of the model are clearly wrong and could be discredited by data. On the other hand, the objective of study of data robustness or data sensitivity, which comes after we have checked that the assumptions of the model cannot be rejected, is to measure how the properties of the model are supported by all the data, and how these basic properties change when some part of data is not taken into account. For example, Figure 2 shows two simple regression models. In the first, the relationship between x and y is clearly supported by all the data, while in the second is based only on two points. Indeed, we would like to know if we are in the first or in the second situation, and, as the complexity of the model increases, the study of this problem becomes more and more important.

As the study of data sensitivity should be based on deleting observations from the model, sample reuse techniques, such as the jackknife and the bootstrap, must be useful. Similarly, cross-validation ideas could also be useful in this stage.

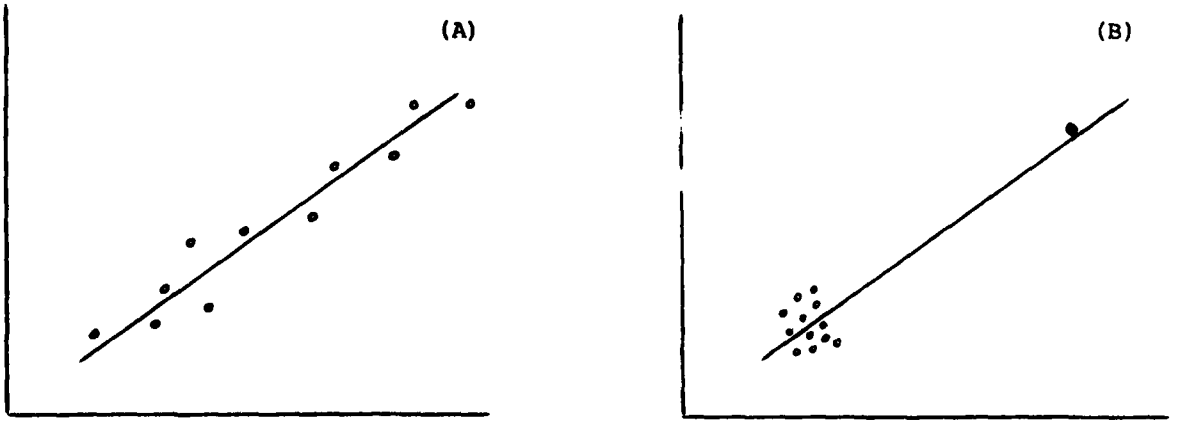


Figure 2

CONCLUSIONS

It has been stated that the unity of statistics lies in its methodology and in its models. These models can be classified in a logical framework which allows a straightforward interpretation of the history of this science and it is helpful to understand why the methodological changes have occurred.

This point is important for teaching and research. For teaching, it stresses the need for a different approach to the traditional presentation in most textbooks. Statistical methodology should be emphasized and the process of "enrichment" of the model through generalization from extrapolative-static models to explicative-dynamic ones should be illustrated with real data. Probability calculus and the theory of stochastic processes should be integrated as the mathematical structure needed to build logical coherent models, and the role of data analysis and descriptive procedures as part of the model building process should be pointed out. From the point of view of research, this analysis has identified areas in which much work needs to be done before a unifed vision of the field can be achieved. Moreover, the analogy betwen the development of models of different types is useful to foresee potentially rewarding lines of investigation.

ACKNOWLEDGEMENTS

I am very grateful to Dennis Cook, David Steinberg, and Jean Wallis for their helpful comments to a preliminary draft of this work.

REFERENCES

[1] Anscombe, F. J. and Tukey, J. W. (1963). "The examination and analysis of residuals". Technometrics, 5, 141-160.

[2] Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. Wiley.

[3] Belsely, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics. Wiley.

[4] Bernoulli, D. (1777). "The most probable choice between several discrepant observations and the formation therefrom of the most likely induction". Reprinted in Biometrika, 48, 1-18, 1961.

[5] Box, G.E.P., Hunter, W. G. and Hunter, J. S. (1978). Statistics for Experimenters. Wiley.

[6] Box, G.E.P. and Jenkins, G. M. (1970). Time Series Analysis, Holden Day.

[7] Box, G.E.P. and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley.

[8] Box, G.E.P. and Cox, D. R. (1964). "An analysis of transformations". JRSS,B, 26, 211-243.

[9] Box, G.E.P. (1983). "The importance of practice in the development of statistics". Technical Report No. 2471, Mathematics Research Center, University of Wisconsin-Madison.

[10] Box, J. F. (1978). Fisher, The Life of a Scientist. New York. Wiley.

[11] Brush, S. G. (1968). "A history of random processes. I. Brownian movement from Brown to Perrin". Archive for the History of Exact Sciences, 5, 1-36.

[12] Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall.

[13] De Finetti, B. (1975). Theory of Probability. Wiley.

[14] Draper, N. R. and Smith H. (1966). Applied Regression Analysis. Wiley (2nd edition, 1980)

[15] Ephremides, A. and Thomas, J. B. (1973). (editor) Random Processes. (Vol. I). Dowden Hutchingon and Ross Inc.

[16] Ephremides, A. (1975). (editor) Random Processes. (Vol. II). Halted Press.

[17] Kendall, D. G. (1966). "Branching processes since 1873". J. London Math. Society, 41, 385-406.

[18] Hannan, E. J. (1970). Multiple Time Series. New York, Wiley.

[19] Harris, T. E. (1963). The Theory of Branching Processes. Berlin.

[20] Jeffreys, H. (1932). "An alternative to the rejection of observations". Proc. Roy. Soc. A, CXXXVII, 78-87.

[21] Karlin, S. and Taylor, H. M. (1975). A First Course in Stochastic Processes. Academic Press.

[22] Lancaster, H. O. (1977). "Development of the notion of statistical dependence", in Studies in the History of Statistics (Vol. II), M. Kendall and R. L. Plackett edit., Mac Millan Publishing Co., Inc.

[23] Maistrov, I. E. (1974). Probability Theory. A Historical Sketch. Academic Press.

[24] Peña, D. (1978). "Modelos con parametros variables en el analisis de series temporales" Questiio, 4, 2, 75-87.

[25] Seal, H. L. (1967). "The historical development of the Gauss linear model". Biometrika, 54, 1-24.

[26] Stigler, S. M. (1975). "Napoleonic Statistics: The Work of Laplace". Biometrika, 62, 3, 503-517.

[27] Thatcher, A. R. (1957). "A note on the early solutions of the problem of the duration of the play". Biometrika, 44, 515-518.

[28] Todhunter, I. (1865). History of the Mathematical Theory of Probability. Cambridge Univ. Press (Reprinted by Chelsea, N.Y., 1949, 1961).

[29] Tukey, J. W. (1960). "A survey of sampling from contaminated distributions" in Contributions to Probability and Statistics, Olkin, I. (edit.), University Press.

[30] Tiao G. C. and Box, G.E.P. (1981). "Modeling Multiple Time Series with Applications". JASA, 76, 376, 802-816.

DP/ed

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2608	2. GOVT ACCESSION NO. AD-A132 936	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ON THE LOGICAL DEVELOPMENT OF STATISTICAL MODELS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Daniel Peña		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE December 1983 13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

statistical models, methodology, history of statistics, robustness

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper presents a classification of statistical models using a simple and logical framework. Some remarks are made about the historical appearance of each type of model and the practical problems that motivated them. It is argued that the current stages of the statistical methodology for model building have arisen in response to the needs of more sophisticated procedures for building dynamic-explicative types of models. Some potentially important topics for future research are included.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)